数据驱动范式下政府开放数据状态及其主体行为状态的关系挖掘*

■ 陈玲[1]　段尧清[1,2]
[1] 华中师范大学信息管理学院　武汉 430079　[2]湖北省数据治理与智能决策研究中心　武汉 430079

摘　要：[目的/意义]在数据驱动范式情境下，揭示政府门户网站开放数据状态及其主体行为状态之间的内部关联，推动政府数据开放效果和进程。[方法/过程]采用爬虫方法抓取上海市政府数据门户网站中各开放数据集，在对各数据集指标进行相关分析的基础上，采用 Stepwise 探索其回归关系，筛选得出关联度较高的变量；进一步对关系显著的变量进行 PLS 回归检验，得出政府开放数据状态与其主体行为状态的内部关联。[结果/结论]在政府数据开放进程中，政府部门的主体行为比数据自身的客体特征对公众主体行为的影响更大。在影响公众评分的因素中，政府开放保密级别的影响因子最大，且具有显著负向影响作用；政府更新频率、政府首次开放时间、数据格式可机读性对公众评分具有显著正向影响作用。

关键词：数据驱动范式　政府开放数据　政府门户网站　数据开放状态
分类号：G203
DOI：10.13266/j.issn.0252－3116.2020.02.002

1　引言

随着数据世界越来越多地反映现实世界，许多传统的管理与决策正在变成数据分析的管理与决策。长期以来，管理学研究一直以模型驱动范式为领域主流。但是在大数据背景下，一些新的挑战正在涌现，使得数据驱动范式的优势不断凸显。数据驱动范式指科学研究第四范式，是针对数据密集型科学，由传统的假设驱动向基于科学数据进行探索的科学方法的转变[1]。主要包括直接发现特定变量关系模式，形成问题解决方案；与模型驱动范式进行补充扩展，形成融合范式[2]。数据驱动范式发现的一类重要关系模式是关联及其扩展形势，如数量关联、时态关联、模式关联等，并广泛应用到许多领域。许多管理决策情境不仅需要关联也需要因果，这在一定程度上推动了数据驱动范式及其应用。

在数据驱动范式情境下，政府部门作为全社会 80% 数据资源的拥有者[3]，其开放数据的效果受到社会各界的重视。研究政府开放数据状态及其主体行为状态之间的关系，有助于推动开放型政府的建设，有助于公众更好地利用政府数据资源，有助于企业创造更多的经济价值和社会价值。目前政府开放数据领域的研究热点可以分为三类：①侧重数据管理的研究，研究热点包括数据安全、数据共享、关联数据、数据挖掘等；②侧重数据开放的研究，研究热点包括开放政府、开放门户、开放模式、开放系统等；③侧重数据服务及应用的研究，研究热点包括智慧城市、公共服务、气候变化、公共卫生等。其中，多数文献侧重研究政府开放数据的现状及开放平台建设等问题。在现状研究方面，部分文献侧重研究国外政府数据开放度较高的国家，在介绍加拿大[3]、美国[4]、英国[5]等国家政府数据开放模式、法律政策、推广机制等方面的基础上，结合我国的具体国情，借鉴其管理和利用的成功经验；部分文献侧重对我国的地方政府数据开放平台进行案例分析，对北京市[6]、上海市[7]、武汉市[8]等开放成果较好的地方政府门户网站进行对比、分析，在梳理其问题与解决对策的基础上，致力于推动我国统一的政府数据开放平台的建成。在开放平台建设方面，分别采用 DEA 数据包络法[9]、BP 神经网络[10]等方法，从数据量、数据内

* 本文系国家社会科学基金重点项目“基于全生命周期的政府开放数据整合利用机制与模式研究”（项目编号：17ATQ006）研究成果之一。
作者简介：陈玲（ORCID：0000－0003－0379－3512），博士研究生，E-mail：2471685835@qq.com；段尧清（ORCID：0000－0002－8991－5842），教授，博士生导师。
收稿日期：2019－05－15　修回日期：2019－08－23　本文起止页码：13－22　本文责任编辑：杜杏叶

容、连通率、下载量、访问量、数据可获性、数据及时性、数据全面性等方面，评价分析了政府开放数据网站的开放效果和建设水平[11-12]。

通过梳理发现，现有文献或者采用定性分析的视角，对开放指标的逻辑属性进行分类、建立标准；或者采用定量分析的视角，对开放指标的关系进行单一属性的研究。本文的创新之处在于，对政府数据集的指标内容在逻辑属性划分的基础上，定量分析不同属性状态，研究指标间的相关关系，并采用了定性与定量相结合的方式。

政府数据开放状态指的是政府部门在开放数据门户网站发布的各条数据集的指标内容，按照指标的逻辑属性可划分为数据状态和主体行为状态[13-14]。数据状态指的是数据自身的属性特征值。主体行为状态按不同的主体属性，又可划分为政府行为状态和公众行为状态[15-16]；政府行为状态指的是政府主体的行为属性值，公众行为状态指的是公众主体的行为属性值。本文的关系挖掘指的是，针对开放数据门户网站发布的各数据集，在逻辑属性分类基础上，测定不同状态下数据集指标内容之间的内部关系[17-19]。

研究首先在数据驱动范式情境下，利用关联挖掘方法，分析政府数据开放不同逻辑属性状态间的相关关系，在此基础上探索其回归关系，以缩减变量空间和组合规模。进而基于关系探测结果，构建政府数据开放状态间的回归路径，检验变量间的回归关系，揭示政府门户网站开放的数据状态、政府主体行为状态、公众行为状态之间的内部关联。

2　研究样本与指标的选取

2.1　研究样本

考虑到目前我国并没有统一的国家政府数据开放门户，而各级地方政府开放数据网站往往采用专门的数据网站形式[20]，网页内容相对统一，因此本文将研究范围限定为地方政府的开放数据网站。通过公开报道和搜索引擎发现已上线开放数据平台的地区，并综合考虑各地开放数据的成熟水平、行政层级和地域分布[21]，同时兼顾是否提供发布数据集的浏览量、下载量、评分等指标以及开放的数据集数量[22]。根据上述条件，本文选择了上海市政府开放数据网站作为研究样本，以此对政府门户网站数据开放状态之间的关系进行研究分析。

2.2　研究变量和指标选取

合理的指标选取是挖掘和分析政府门户网站数据开放状态关系的重要前提。本文构建的开放状态类别和研究指标是基于上海市政府开放数据网站的实际运行情况，具体如表 1 所示。其中，数据状态包括数据格式多样性和数据格式可机读性；政府行为状态包括政府开放保密级别、政府更新频率、政府首次开放时间；公众行为状态包括公众浏览、公众下载、公众评分。

表 1　政府数据开放状态的测度指标

状态类别	研究指标	指标内容	文献来源
数据状态	数据格式多样性	Berners-Lee 的五星评估等级	文献[4][7][21]
	数据格式可机读性		文献[7][11][20]
政府行为状态	政府开放保密级别	数据的公开属性：具有数据开放授权协议、免费获取、自由利用	文献[5][17-20]
	政府更新频率	数据的更新频率：按年、月、周、日等	文献[6-8][10][20]
	政府首次开放时间	政府开放数据的首发日期，首次发布时间	文献[7][22][23]
公众行为状态	公众浏览	政府开放数据的浏览量统计	文献[9][12]
	公众下载	政府开放数据的下载量统计	文献[9][12]
	公众评分	政府开放数据的公众评分	文献[10]

2.3　数据采集

针对网站的不同分类主题，研究主要采用爬虫软件，并结合人工观察的方式抓取了以上指标数据。数据采集时间截止到 2018 年 12 月 31 日 10 点，共采集数据集 1 233 条，所有数据均来自于上海市政府开放数据门户网站，具有真实可靠性。各主题分类的数据集数量如表 2 所示：

表 2　各主题数据集统计

主题分类	城市建设	道路交通	公共安全	机构团体	教育科技	经济建设	民生服务	社会发展	卫生健康	文化休闲	信用服务	资源环境	总计
数据集量（条）	136	128	30	49	134	338	90	49	131	向69	11	68	1 233

3 数据处理与检验

3.1 数据清洗与数值化

本文研究的是政府门户网站数据开放状态之间的关系，因此，在进行数据清洗时，分别剔除无数据格式、无公开属性、无更新频率、无首发日期、无浏览量、无下载量、无评分的数据集共计 605 条，最终得到有效数据集共计 628 条。

数据集各项指标内容的数值化过程见表 3。每条数据集包含格式多样性、格式可机读性、公开属性、更新频率、首发日期、浏览量、下载量、评分等级 8 项指标，根据各自的属性和特征值，将其数值化[23－25]。

其中，公开属性包括普通公开和特定公开，依次将其数值化为 1 和 2。更新频率包括一次性、每五年/每十年、每年、每半年、每季度、每月、每两周、每周、按需、实时/即时，依次将其数值化为 1－10。首发日期按年份包括 2012 年－2018 年，数值化为其发布年份到 2018 年的年份差，依次将其数值化为 1－7。数据格式包括 PDF、RAR、ZIP、XLS \ XLSX、DOC \ DOCX、XML、CSV 等，格式可机读性按 Berners-Lee 的五星评估等级将其数值化为 1－3。格式多样性按其开放数据的格式类型的数量多少，将其进行数值化。浏览量和下载量为数值型数据，分别为该条政府数据的累计浏览量和累计下载量。评分等级包括★、★★、★★★、★★★★、★★★★★，依次将其数值化为 1－5。

3.2 数据描述统计与标准化处理

本研究对样本数据进行结构变量统计，得出相应的极小值、极大值、均值、标准差等统计量，具体见表 4。从表 4 可以看出，评分等级、公开属性、更新频率、首发日期、格式可机读性、格式多样性的标准差都小于 2，这表明各变量之间的差异度和离散度均比较小。而浏览量和下载量的标准差分别为 3 436.144、2 751.472，这是因为评分等级和公开属性等 6 个研究变量为量表级数据，而浏览量和下载量为数值型数据。

据此，研究对样本数据进行"Z"标准化处理，将每一变量值与其平均值之差除以该变量的标准差。通过这种标准化处理可以消除量纲和数量级的影响，去除数据的单位限制，将其转化为无量纲的纯数值。其数据转换函数为：

$$X^* = (x - \mu)/\sigma$$

其中，X^* 为标准化后的变量值，x 为实际观测值，μ 为所有样本数据的均值，σ 为所有样本数据的标准差。标准化结果（部分）见表 5。

表 3　政府开放数据指标的数值化

测度指标内容	指标数值化
公开属性	普通公开→1
	特定公开→2
更新频率	一次性→1
	每五年/每十年→2
	每年→3
	每半年→4
	每季度→5
	每月→6
	每两周→7
	每周→8
	按需→9
	实时/即时→10
首发日期	2018 年→1
	2017 年→2
	2016 年→3
	2015 年→4
	2014 年→5
	2013 年→6
	2012 年→7
格式可机读性	PDF、RAR、ZIP→1
	XLS\XLSX、DOC\DOCX→2
	XML、CSV→3
格式多样性	格式类型的数量值
浏览量	浏览量的数量值
下载量	下载量的数量值
评分等级	★→1
	★★→2
	★★★→3
	★★★★→4
	★★★★★→5

表 4　政府开放数据的描述统计量

描述统计量	极小值	极大值	均值	标准差
公众评分	1	5	3.76	.919
公众浏览	724	50 911	2 659.09	3 436.144
公众下载	48	40 287	700.21	2 751.472
政府开放保密级别	1	2	1.19	.391
政府更新频率	2	10	4.77	1.324
政府首次开放时间	1	6	3.46	1.265
数据格式可机读性	2	3	2.81	.391
数据格式多样性	1	3	1.89	.416

表 5　政府开放数据的 Z 标准化结果(部分)

Z 公众评分	Z 公众浏览	Z 公众下载	Z 政府开放保密级别	Z 政府更新频率	Z 政府首次开放时间	Z 数据格式可机度性	Z 数据格式多样性
-3.009 17	-0.520 38	-0.173 44	-0.480 63	-1.683 86	-1.942 99	-2.077 29	0.260 13
-3.009 17	-0.560 25	-0.195 97	-0.480 63	-1.683 86	-1.942 99	-2.077 29	0.260 13
-3.009 17	-0.502 92	-0.209 78	2.077 29	-1.683 86	-1.152 71	-2.077 29	0.260 13
-3.009 17	-0.550 06	-0.204 33	-0.480 63	-1.683 86	-1.942 99	-2.077 29	0.260 13
-3.009 17	-0.470 32	-0.210 51	2.077 29	-1.683 86	-1.152 71	-2.077 29	0.260 13
-3.009 17	0.273 25	0.154 75	2.077 29	-2.439 20	-0.362 42	-2.077 29	-2.142 27
-3.009 17	0.140 83	-0.043 32	2.077 29	-2.439 20	-0.362 42	-2.077 29	-2.142 27
-3.009 17	-0.428 41	-0.225 41	-0.480 63	-2.439 20	-1.152 71	-2.077 29	-2.142 27
-3.009 17	-0.406 88	-0.201 42	-0.480 63	-2.439 20	-1.942 99	-2.077 29	0.260 13
-3.009 17	-0.111 78	-0.100 75	2.077 29	-2.439 20	-0.362 42	-2.077 29	-2.142 27

3.3　CMV 检验

针对共同方法偏差的问题，本研究在录入数据库、整理数据后进行了分析控制。采用 Harman 单因素检验，检验结果见表 6。由表 6 可知，因素分析得到 8 个因子，总和大于 1；且第一个因子总和为 2.613，方差百分比为 32.668%，低于临界值 40%。因此，共同方法偏差对本研究影响不大。

表 6　政府开放数据的解释总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	2.613	32.668	32.668	2.613	32.668	32.668
2	1.958	24.480	57.148	1.958	24.480	57.148
3	1.349	16.861	74.009	1.349	16.861	74.009
4	.778	9.729	83.739			
5	.622	7.780	91.519			
6	.393	4.910	96.429			
7	.271	3.383	99.812			
8	.015	.188	100.000			

4　政府数据开放状态之间的关系探索

4.1　政府数据开放状态之间的相关分析

在对样本数据进行清洗、数值化及标准化处理、CMV 检验之后，对研究变量及其测度指标进行了相关分析，结果如表 7 所示。由表 7 可知：公众评分与数据状态、政府行为状态各变量之间的关联度较高，但与公众浏览、公众下载之间的相关性不显著。此外，公众浏览、公众下载之间具有显著相关性，但与其他各研究变量之间不具有显著相关关系。

表 7　政府数据开放状态之间的相关分析

相关值	公众评分	公众浏览	公众下载	政府开放保密级别	政府更新频率	政府首次开放时间	数据格式可机度性	数据格式多样性
公众评分	1	.061	.038	-.525 **	.509 **	.343 **	.454 **	.175 **
公众浏览	.061	1	.984 **	-.063	.070	.108 **	.024	-.005
公众下载	.038	.984 **	1	-.052	.058	.067	.024	.004
政府开放保密级别	-.525 **	-.063	-.052	1	-.287 **	-.355 **	-.228 **	-.012
政府更新频率	.509 **	.070	.058	-.287 **	1	.190 **	.410 **	.178 **
政府首次开放时间	.343 **	.108 **	.067	-.355 **	.190 **	1	.168 **	.010
数据格式可机度性	.454 **	.024	.024	-.228 **	.410 **	.168 **	1	.659 **
数据格式多样性	.175 **	-.005	.004	-.012	.178 **	.010	.659 **	1

**:在 0.01 上显著

4.2　政府数据开放状态之间的回归分析

采用 Stepwise 进一步探索公众行为状态与数据状态、政府行为状态各变量之间的回归关系。为避免回归模型中可能存在的多重共线性问题，对所有变量进行中心化处理。首先，对 p 个回归自变量 $X_1, X_2, \cdots, X_p$ 分别同因变量 Y 建立一元回归模型：$Y = \beta_0 + \beta_i X_i + \varepsilon, i = 1, \cdots, p$。计算变量 X_i 相应的回归系数的 F 检验统计量的值，记为 $F_1^{(1)}, \cdots, F_p^{(1)}$，取其中的最大值 $F_{i1}^{(1)}$，即 $F_{i1}^{(1)} = \max\{F_1^{(1)}, \cdots, F_p^{(1)}\}$。对给定的显著性水平 α，记相应的临界值为 $F^{(1)}$，$F_{i1}^{(1)} \geq F^{(1)}$，则将变量 X_{i1} 引入回归模型，记 I_1 为选入变量指标集合。其次，建立因变量 Y 与自变量子集 $\{X_{i1}, X_1\}, \cdots, \{X_{i1},$

$X_{i1-1}\}$，$\{X_{i1},X_{i1+1}\}$，…，$\{X_{i1},X_p\}$的二元回归模型，共有p-1个。计算变量的回归系数F检验的统计量值，记为$F_K^{(2)}(K\neq I_1)$，选其中最大者，记为$F_{i2}^{(2)}$，对应自变量脚标记为i_2，即$F_{i2}^{(2)}=\max\{F_1^{(2)},\cdots,F_{i1-1}^{(2)},F_{i1+1}^{(2)},\cdots,F_p^{(2)}\}$。对给定的显著性水平α，记相应的临界值为$F^{(2)}$，$F_{i2}^{(1)}\geqslant F^{(2)}$，则变量$X_{i2}$引入回归模型。否则，终止变量引入过程。最后，考虑因变量对变量子集$\{X_{i1},X_{i2},X_k\}$的回归重复上述步骤，每次从未引入回归模型的自变量中选取一个，直到没有变量引入为止。

逐步回归分析结果显示：公众行为状态中，公众浏览、公众下载与其他研究变量之间的回归关系不显著，公众评分与政府开放保密级别、政府更新频率、政府首次开放时间、数据格式可机读性之间具有显著回归关系。故此，研究主要对公众评分与政府行为状态、数据状态各变量之间的回归分析结果进行了进一步解释说明，具体见表8-表11。

由表8可知，逐步层级回归中公众评分为被解释变量，最后得到的解释变量集包括政府开放保密级别、政府更新频率、数据格式可机读性、政府首次开放时间。共有4个假设模型，模型1中仅纳入政府开放保密级别；模型2在模型1的基础上纳入政府更新频率；模型3在模型2的基础上纳入数据格式可机读性；模型4在模型3的基础上纳入政府首次开放时间。

由表9和表10可知，4个假设模型的Sig系数均在0.001之下，说明政府开放保密级别、政府更新频率、数据格式可机读性、政府首次开放时间这几个解释变量对公众评分的回归作用均显著。但假设模型4的R方值和调整R方值相较于其他三个模型均最大，说明其对公众评分的解释效果最好。

由表11中的共线性统计量可知，所有变量的容差均在0.1之上，且方差膨胀因子(VIF)均远小于10，因此回归中不存在多重共线性问题。进一步观察表11中的标准化和非标准化系数可知，政府开放保密级别对公众评分具有显著负向影响作用，政府更新频率、数据格式可机读性、政府首次开放时间对公众评分具有显著正向影响作用。

表8　输入/移去的变量[a]

模型	输入的变量	移去的变量	方法
1	政府开放保密级别	-	步进(准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100)
2	政府更新频率	-	步进(准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100)
3	数据格式可机读性	-	步进(准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100)
4	政府首次开放时间	-	步进(准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100)

a. 因变量：公众评分

表9　模型汇总[e]

模型	R	R方	调整R方	标准估计误差	更改统计量			Durbin-Watson
					R方更改	F更改	Sig. F更改	
1	.525[a]	.276	.274	.783	.276	238.044	.000	
2	.644[b]	.415	.413	.704	.140	149.459	.000	
3	.681[c]	.464	.461	.674	.049	56.732	.000	
4	.691[d]	.478	.474	.666	.014	16.379	.000	.373

a. 预测变量：(常量)，政府开放保密级别。b. 预测变量：(常量)，政府开放保密级别，政府更新频率。c. 预测变量：(常量)，政府开放保密级别，政府更新频率，数据格式可机读性。d. 预测变量：(常量)，政府开放保密级别，政府更新频率，数据格式可机读性，政府首次开放时间。e. 因变量：公众评分

5　政府数据开放状态之间的关系检验

5.1　PLS回归路径构建

研究对政府门户网站数据开放状态之间的回归路径进行了PLS回归检验。PLS检验中研究变量与观测指标间的关系，通常可用3个矩阵方程式表达：

$$\eta=B\eta+\Gamma\xi+\zeta$$

$$X=\Lambda_x\xi+\delta$$

$$Y==\Lambda_y\xi+\varepsilon$$

其中，Λ_x为外生观测变量与外生潜变量直接的关系，是外生观测变量在外生潜变量上的因子载荷矩阵；Λ_y为内生观测变量与内生潜变量之间的关系，是内生观测变量在内生潜变量上的因子载荷矩阵；B为路径系数，表示内生潜变量间的关系；Γ为路径系数，表示外生潜变量对内生潜变量的影响；ζ为结构方程的残差项，反映了在方程中未能被解释的部分。

表 10 Anova[a]

模型		平方和	df	均方	F	Sig.
1	回归	172.739	1	172.739	238.044	.000b
	残差	454.261	626	.726		
	总计	627.000	627			
2	回归	260.404	2	130.202	221.978	.000c
	残差	366.596	625	.587		
	总计	627.000	627			
3	回归	290.957	3	96.986	180.093	.000d
	残差	336.043	624	.539		
	总计	627.000	627			
4	回归	299.565	4	74.891	142.493	.000e
	残差	327.435	623	.526		
	总计	627.000	627			

a. 因变量：公众评分；b. 预测变量：(常量)，政府开放保密级别；c. 预测变量：(常量)，政府开放保密级别，政府更新频率；d. 预测变量：(常量)，政府开放保密级别，政府更新频率，数据格式可机读性；e. 预测变量：(常量)，政府开放保密级别，政府更新频率，数据格式可机读性，政府首次开放时间

表 11 系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准误差	试用版			容差	VIF
1	(常量)	1.023E-013	.034		.000	1.000		
	政府开放保密级别	-.525	.034	-.525	-15.429	.000	1.000	1.000
2	(常量)	1.027E-013	.031		.000	1.000		
	政府开放保密级别	-.413	.032	-.413	-12.935	.000	.918	1.090
	政府更新频率	.390	.032	.390	12.225	.000	.918	1.090
3	(常量)	1.024E-013	.029		.000	1.000		
	政府开放保密级别	-.384	.031	-.384	-12.441	.000	.903	1.107
	政府更新频率	.299	.033	.299	9.073	.000	.792	1.262
	数据格式可机度性	.244	.032	.244	7.532	.000	.819	1.222
4	(常量)	1.023E-013	.029		.000	1.000		
	政府开放保密级别	-.343	.032	-.343	-10.712	.000	.816	1.226
	政府更新频率	.290	.033	.290	8.884	.000	.789	1.268
	数据格式可机度性	.236	.032	.236	7.350	.000	.815	1.227
	政府首次开放时间	.126	.031	.126	4.047	.000	.862	1.160

a. 因变量：公众评分

PLS 结果显示，公众浏览被政府行为状态及数据状态的解释程度仅为 0.015，公众下载被政府行为状态及数据状态的解释程度仅为 0.007，公众评分被政府行为状态及数据状态的解释程度为 0.478，总体公众行为状态被政府行为状态及数据状态的解释程度为 0.470，小于 0.478。据此本研究仅对公众评分的 PLS 检验结果进行了进一步展示。检验时发现，数据格式多样性对公众评分的显著性 T 值仅为 1.455，据此将其剔除。最终构建的路径模型图见图 1。

对回归路径模型进行信度、效度检测，整体检测结果见表 12。样本数据为政府网站爬取的客观数据，各

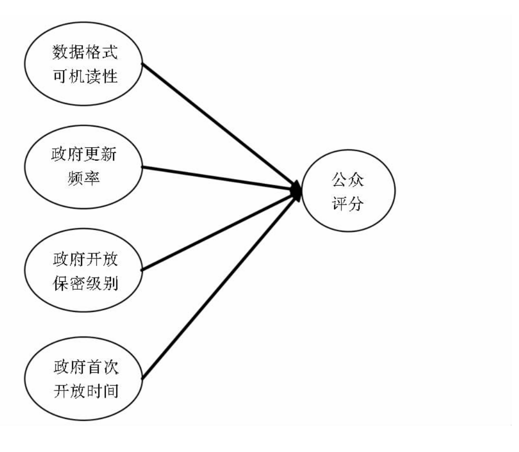


图 1 回归路径模型

变量的Cronbachs Alpha系数和C.R.系数均在0.7以上,说明具有良好的内部一致性和稳定性。且AVE值均在0.5之上,说明收敛效度较好,具有准确性。此外,公众评分的R方为0.478,说明其被解释程度较高。

表12　整体检测结果

检测变量	AVE	Composite Reliability	R Square	Cronbachs Alpha	Communality	Redundancy
公众评分	1.000 000	1.000 000	0.477 775	1.000 000	1.000 000	0.242 544
政府开放保密级别	1.000 000	1.000 000		1.000 000	1.000 000	
政府更新频率	1.000 000	1.000 000		1.000 000	1.000 000	
政府首次开放时间	1.000 000	1.000 000		1.000 000	1.000 000	
数据格式可机读性	1.000 000	1.000 000		1.000 000	1.000 000	

5.2　PLS回归路径显著性检测

回归路径的显著性检测结果见表13。由表13可知,政府首次开放时间、政府开放保密级别、政府更新频率、数据格式可机读性到公众评分这4条回归路径的显著性T值均大于1.96,说明其对公众评分均具有显著影响作用。

表13　回归路径检测结果

检测变量	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	Standard Error (STERR)	T Statistics (|O/STERR|)
政府开放保密级别 → 公众评分	-0.343 342	-0.343 027	0.035 436	0.035 436	9.689 101
政府更新频率 → 公众评分	0.289 624	0.289 697	0.028 723	0.028 723	10.083 308
政府首次开放时间 → 公众评分	0.126 215	0.126 839	0.024 756	0.024 756	5.098 408
数据格式可机读性 → 公众评分	0.235 662	0.234 257	0.028 253	0.028 253	8.341 236

5.3　PLS回归路径系数检测

在对回归路径进行显著性检测之后,进一步检验模型中各条回归路径的系数,具体检测结果如图2和表14所示。由图2和表14可以看出,公众评分的R方为0.478,说明回归路径模型的拟合效果较好。其中,政府首次开放时间→公众评分、政府开放保密级别→公众评分、政府更新频率→公众评分、数据格式可机读性→公众评分这4条路径的系数值分别为0.126、-0.343、0.290、0.236。综上所述,我们可以认为:政府开放保密级别对公众评分的影响程度最大,且具有显著负向影响作用;政府首次开放时间、政府更新频率、数据格式可机读性对公众评分具有显著正向影响作用。这与前文逐步回归探测结果具有一致性。

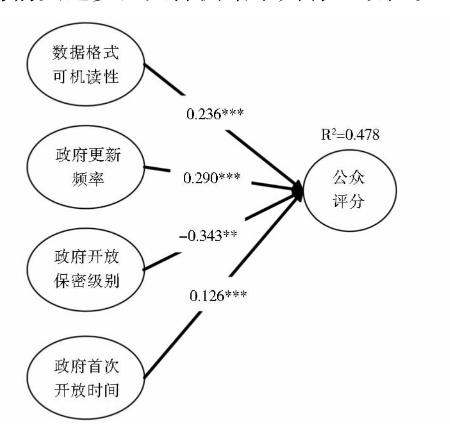


图2　回归路径系数

表14　相关系数矩阵

变量	公众评分	政府开放保密级别	政府更新频率	政府首次开放时间	数据格式可机读性
公众评分	1.000 000				
政府开放保密级别	-0.524 881	1.000 000			
政府更新频率	0.508 731	-0.286 754	1.000 000		
政府首次开放时间	0.342 839	-0.355 016	0.190 371	1.000 000	
数据格式可机读性	0.453 826	-0.227 783	0.410 013	0.168 021	1.000 000

6　结论与建议

6.1　小结

本文的创新之处在于:①对政府数据集的指标内容在逻辑属性划分的基础上,定量分析不同属性状态间、研究指标的相关关系,采用了定性与定量相结合的方式。②在数据驱动范式情境下,利用关联挖掘方法,分析政府数据开放不同逻辑属性状态间的相关关系。③在缩减变量空间和组合规模的基础上,检验变量间的回归关系,揭示数据状态、政府主体行为状态、公众

行为状态之间的内部关联。

在对政府数据开放状态进行相关分析、stepwise 回归探索和 PLS 回归检验中发现：公众浏览、公众下载与其他开放状态之间的回归关系不显著，且被其他开放状态解释的程度很低；而公众评分与其他开放状态之间具有显著回归关系，且被解释程度较高，故此进行了进一步展示与解释说明。根据上文分析结果，对比陈列出政府数据开放状态之间关系的探索、验证结果，具体如表 15 所示：

表 15 政府数据开放状态之间关系的探索、验证结果

公众行为状态变量	研究方法	政府行为状态变量			数据状态变量	
		政府开放保密级别	政府更新频率	政府首次开放时间	数据格式可机读性	数据格式多样性
公众评分	相关关系	-	+	+	+	+
	stepwise 回归	-	+	+	+	不显著
	PLS 回归检验	-	+	+	+	不显著
公众浏览	相关关系	不显著	不显著	+	不显著	不显著
	stepwise 回归	不显著	不显著	+	不显著	不显著
	PLS 回归检验	不显著	不显著	+	不显著	不显著
公众下载	相关关系	不显著	不显著	不显著	不显著	不显著
	stepwise 回归	不显著	不显著	不显著	不显著	不显著
	PLS 回归检验	不显著	不显著	不显著	不显著	不显著

由表 15 和上文分析结果可知：

（1）公众浏览、公众下载之间具有显著相关性，但与其他开放状态之间的关联度较低，且公众浏览、公众下载被政府行为状态、数据状态解释的程度很低。

（2）公众评分与政府行为状态、数据状态各研究变量之间的关联度较高。具体表现在：①在政府数据开放进程中，政府部门的主体行为比数据自身的客体特征，对公众评分行为的影响更大。在影响公众评分的因素中，政府开放保密级别的影响程度 > 政府更新频率的影响程度 > 数据格式可机读性的影响程度 > 政府首次开放时间的影响程度。②政府开放保密级别对公众评分的影响最为显著，且具有显著负向影响作用。该研究变量的观测指标——公开属性主要包括普通公开和特定公开。受政府数据保密性的影响，特定公开属性的政府数据，其保密性较高，获得的公众好感度相对较低。③政府更新频率对政府开放数据评分具有显著正向影响作用。此次采集的政府数据的更新频率分为一次性、每五年/每十年、每年、每半年、每季度、每月、每两周、每周、按需、实时/即时。更新频率越高的数据，更具有时效性，更容易获得公众的好感度。④数据格式可机读性对政府开放数据评分具有显著正向影响作用。XML 等具有结构性的数据和 CSV 等以纯文本形式存储的数据，更容易得到公众的好感度；XLS\XLSX、DOC\DOCX 等文档和表格数据，也较容易获得公众好评；而 RAR、ZIP 等压缩文件格式，由于使用的不便性，不易得到公众的好评。⑤政府首次开放时间对政府开放数据评分具有显著正向影响作用。此次研究爬取的政府网站数据，是基于某一个静态的时间节点，越早发布的政府数据，积累的公众好感度越高；而首发日期较晚的政府数据，积累的公众好感度则相对较低。

6.2 建议

结合上述分析，本研究认为可以从以下几个方面推动政府开放数据效果和进程：

（1）公众浏览行为、公众下载行为虽然与政府行为状态、数据状态之间的关联度较低，但浏览与下载之间具有显著的强相关关系。因此在政府数据开放进程中，要坚持“以公众为中心”并时刻关注公众体验[26]，要把为公众提供所需数据作为建设政府开放门户网站的中心目标；要扩大政府开放数据的社会影响，使政府数据开放更贴近公众生活，解决与公众的生活息息相关的问题。

（2）公众评分行为受政府行为状态、数据状态各变量之间的影响较大，为此政府部门应开放更多的高价值数据，提高政府数据的经济价值与社会价值。包括：①在遵循保密性原则和重视个人隐私的前提下，提升政府数据的开放属性，降低保密级别，做到真正的开放；②提高数据更新频率，尽量做到实时更新和按时更新，加强开放数据的稳定性和时效性；③加大开放数据的可机读性[27]，提高政府数据的易用性水平，易于计算机自动读取和处理，方便用户获取和利用；④尽早开放和发布各主题类型的数据，扩大开放数据的范围。

（3）针对此次研究过程中发现的问题，建议进一步推进国家层面政府数据开放门户网站的建成，同时

尽快统一地方政府和各部门政府的数据平台建设标准[28]。我国政府数据的开放与利用还未发展成为全国范围内的行动,且地方层面的政府部门在开放数据进程中没有相关政策法规的指导。为此,国家应尽快出台并完善政府数据开放相关的法律法规,建立开放数据标准规范,保障政府数据开放工作有序展开;地方政府应积极参与到开放平台建设进程中,为建立统一的、集成的全国性的数据平台打好基础。

6.3 不足与展望

我国目前没有国家级层面的政府数据开放门户网站,且地方开放平台建设水平、标准不统一,各地方门户的开放数据集数量、开放属性、元数据开放指标等参差不齐,使得全样本分析与对比分析具有一定的困难与障碍。本文综合考虑地方层面的平台开放现状与理论研究价值,仅选取了上海市政府开放数据网站为研究样本,基于理论基础,从上海市门户网站的实际运行情况出发,选取了研究变量与指标。针对本文产生的不足之处,将在今后的研究中考虑如何解决全样本分析与各地方层面的对比分析,且在研究变量与测度指标的选取方面展开进一步深化和探讨。

参考文献:

[1] 刘雨农, 是沁. 数据驱动范式下的人文社科知识服务创新研究[J]. 图书与情报, 2019(1):24－30.
[2] 邓仲华, 李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作, 2013, 34(4):19－23.
[3] 杨菲菲. 国外政府数据开放门户网站建设研究[D]. 河北:燕山大学,2016.
[4] 侯人华, 徐少同. 美国政府开放数据的管理和利用分析——以www.data.gov为例[J]. 图书情报工作, 2011, 55(4):119－122.
[5] 李重照, 黄璜. 英国政府数据治理的政策与治理结构[J]. 电子政务, 2019(1):25－36.
[6] 黄思棉, 张燕华. 当前中国政府数据开放平台建设存在的问题与对策研究——以北京、上海政府数据开放网站为例[J]. 中国管理信息化, 2015(14):175－177.
[7] 顾铁军, 夏媛, 徐柯伟. 上海市政府从信息公开走向数据开放的可持续发展探究——基于49家政府部门网站和上海政府数据服务网的实践调研[J]. 电子政务, 2015(9):14－21.
[8] 陈涛, 李明阳. 数据开放平台建设策略研究——以武汉市政府数据开放平台建设为例[J]. 电子政务, 2015(7):46－52.
[9] 马海群, 王今. 基于DEA的政府开放数据网站效率评价[J]. 数字图书馆论坛, 2016(6):2－7.
[10] 邹纯龙, 马海群. 基于神经网络的政府开放数据网站评价研究以美国20个政府开放数据网站为例[J]. 现代情报, 2016, 36(9):16－21.
[11] 郑磊, 熊久阳. 中国地方政府开放数据研究:技术与法律特性[J]. 公共行政评论, 2017, 10(1):53－73.
[12] 段尧清, 邱雪婷, 何思奇. 主题与区域视角下我国城市政府开放数据利用现状分析[J]. 图书情报工作, 2018,62(20):65－76.
[13] 杨永清, 张金隆, 满青珊, 等. 移动互联网用户采纳研究——基于感知利益、成本和风险视角[J]. 情报杂志, 2012, 31(1):200－206.
[14] 刘文奇. 中国公共数据库数据质量控制模型体系及实证[J]. 中国科学:信息科学, 2014, 44(7):836－856.
[15] 李明帅, 管桦. 基于内容分类的政务微博关注度分析——以四川省政府政务微博为例[J]. 情报探索, 2014(12):12－15.
[16] 赵丽棉, 黄基廷. 中国城镇居民消费支出的多元非线性回归模型研究[J]. 数学的实践与认识, 2011, 41(10):20－25.
[17] SOLAR M, DANIELS F, LOPEZ R, et al. A model to guide the open government data implementation in public agencies[J]. Journal of universal computerence, 2014, 20(11):1564－1582.
[18] ZELETI F, OJO A, CURRY E. Exploring the economic value of open government data[J]. Government information quarterly, 2016, 33(3):535－551.
[19] CHARALABIDIS Y, ALEXOPOULOS C, LOUKIS E. A taxonomy of open government data research areas and topics[J]. Journal of organizational computing & electronic commerce, 2016, 26(1):41－63.
[20] 曹雨佳. 政府开放数据生存状态:来自我国19个地方政府的调查报告[J]. 图书情报工作,2016,60(14):94－101.
[21] 赵蓉英,梁志森,段培培. 英国政府数据开放共享的元数据标准——对Data.gov.uk的调研与启示[J]. 图书情报工作,2016,60(18):1－9.
[22] JING D, WEN L. Study on the government strategy of transformation from information publishing to data opening[M]. USA: 2017 IEEE 2nd International Conference on Big Data Analysis,2017.
[23] THOHARI A H, SUHARDI S. Requirement engineering for open government information network development to support digital startup in Cimahi city Indonesia[C]// Proceedings of International Conference on Information Technology Systems and Innovation. Piscataway: IEEE,2016.
[24] 黄如花, 温芳芳, 黄雯. 我国政府数据开放共享政策体系构建[J]. 图书情报工作, 2018, 62(9): 5－13.
[25] 孙璐, 李广建. 政府开放数据应用分析模型构建研究[J]. 图书情报工作, 2017,61(3):97－108.
[26] 周文泓. 新西兰政府数据开放的特点及其启示[J]. 图书情报工作, 2017,61(23),76－82.
[27] 韦忻伶, 安小米, 李雪梅, et al. 开放政府数据评估体系述评:特点分析[J]. 图书情报工作, 2017,61(18):119－127.
[28] 夏义堃. 论政府数据开放风险与风险管理[J]. 情报学报, 2017(1):22－31.

作者贡献说明:

陈玲:负责大纲拟定、资料收集、数据分析与论文初稿撰写;

段尧清:负责论文选题、全文深度修改。

Relation Mining Between Government Open Data State and the Subject Behavior State in the Data Driven Paradigm

Chen Ling[1] Duan Yaoqing[1,2]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] Hubei Data Governance and Intelligent Decision Research Center, Wuhan 430079

Abstract: [Purpose/significance] In the context of data-driven paradigm, this paper reveals the internal relationship between the open data state of government portal website and the behavior state of its main body, and promotes the effect and process of government data opening. [Method/process] This paper used the crawler method to grab the open data sets in the Shanghai government data portal, and then did correlation analysis and Stepwise regression analysis on the index variables of each data set in turn, and screened out the variables with high correlation degree. At the same time, we further carried out PLS regression test on the variables with significant relationships, and ultimately drew the internal relationship between the state of government open data and the state of its main body's behavior. [Result/conclusion] In the process of government data opening, the subject behavior of government departments has a greater impact on public subject behavior than the object characteristics of data itself. Among the factors affecting the public rating, the sequence from high to low is: government openness and secrecy, machine readability of Data Format, the first opening time of government, the timeliness of government openness. The level of government openness and secrecy has a significant negative impact on the score of government open data; the first opening time of government, the timeliness of government openness, and the machine readability of data format have a significant positive impact on the public score.

Keywords: data driven paradigm　government open data　government portals　data open state